



## **CO4760: Exploratory Data Analysis**

### **Assignment 2**

**Date issued:** 14/02/2020

**Hand in Date:** 10/04/2020

**Presentation Hand in Date:** 08/05/2020

**Presentation Date:** during May's exam period (tba)

**Assignment 2 accounts for 60% of the total score.**

### **IMPORTANT**

- As work is submitted on-line, **the deadline is midnight on the hand in date.**
- **Read the marking scheme carefully.**
- **This is an individual project** and no group work is permitted.

### **Assignment Purpose and Overview**

The purpose of this assignment is to help students better understand the exploratory data analysis techniques taught in class sessions by utilizing them on a real dataset. You will be provided with a real dataset for which you will develop your own R code and perform the necessary techniques for descriptive and exploratory data analysis; you will conclude your analysis through critical assessment of the discovered statistical results.

In particular, in this assignment, you will

- perform inferential data analysis on the data set using R
  - perform hypothesis testing (e.g. t-test, correlation test, etc.) where applicable and appropriate;
  - evaluate regression modelling concepts based on the characteristics of the data;
  - perform the appropriate diagnostics and transformations so as to retrieve the most useful model;
- make judgments and critically assess the results you will obtain to conclude the data analysis.

### **Description of the datasets**

You will use the dataset you used in Assignment 1, which you are already familiar with.

## Deliverables

You will be required to submit **one** report document in MS Word format (.docx) and **one** presentation in MS PowerPoint format (.pptx) on the corresponding Hand in Date.

### Report

The report will be a **continuation of the report you submitted in Assignment 1**, and it must contain an Executive Summary, a table of contents, an Introduction in which you will describe the dataset, aim and objectives of your study, one or more sections describing your work (descriptive analysis, summaries, graphical representations, inferential analysis, testing, regression, diagnostics), a Conclusions section and an Annex displaying your R code (from both Assignments) with all commands commented.

The structure of the report must include the following:

Part	Description	Range
Executive Summary	Description of the dataset the problem and the aim and objectives of your study. Discussion of the methods that were employed in the analysis and key results that result from the analysis	0-9
Descriptive statistical analysis	Description of the descriptive statistical analysis performed including appropriate visualizations.	-
Inferential statistical analysis	Description of the inferential statistical analysis performed including appropriate visualizations.	0-25
R code (Annex)	The R code implementation including appropriate comments	0-11

In addressing the work within each section, factors such as simplicity, quality and appropriateness of comments, and quality and completeness of the design will be considered.

### Presentation

The presentation will present the dataset, aim and results of your study, and it will be based upon **both** Assignments. It can be divided in three parts: Introduction (description of the dataset and aim), Descriptive analysis (based on Assignment 1), Inferential analysis (based on Assignment 2).

The structure of the presentation must include the following:

Part	Description	Range
Introduction	Description of the dataset, aim and objectives of your study.	0-1
Descriptive statistical analysis	Presentation of the descriptive analysis methods employed and results.	0-5
Inferential statistical analysis	Presentation of the inferential analysis methods employed and results.	0-5

Conclusions	Conclusions of your study, key results from your analysis.	0-2
Questions	Answering of questions made by the class after the presentation.	0-2

On the presentation date you will have **10 minutes** to present the results of your analysis to the class and 2-3 minutes for questions.

### Grading Criteria

A total of **60 marks** will be awarded based on the criteria described below.

Description	Range	Break down
Executive Summary  (additions to the existing executive summary of Assignment 1)	0-9	<p>0-2 Report</p> <ul style="list-style-type: none"> <li>0: no attempt</li> <li>1: poorly written and presented report</li> <li>2: high quality executive report</li> </ul> <p>0-2 Description of the problem</p> <ul style="list-style-type: none"> <li>0: no attempt</li> <li>1: high level description of the problem and the aim of the study</li> <li>2: detailed description of the problem, the aim and objectives of the study</li> </ul> <p>0-2 Methodology</p> <ul style="list-style-type: none"> <li>0: no attempt or fails to understand the needs of the problem</li> <li>1: high level description of the methods that can be used for the analysis</li> <li>2: evidence of a methodology that will be utilized for the study</li> </ul> <p>0-3: Results</p> <ul style="list-style-type: none"> <li>0: no attempt</li> <li>1: description of the key results</li> <li>2: adequate conclusions are drawn, shows understanding of the statistical results</li> <li>3: draws all the conclusions that result from the analysis, shows understanding of the statistical results and critically assesses and interprets the results of the analysis</li> </ul>
Inferential statistical analysis	0-25	<u>0-9: Hypothesis testing</u>

Description	Range	Break down
		<p>0-3: t-tests / 0-3: correlation tests / 0-3: normality tests</p> <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 1: performs limited tests that are inadequate to assess key hypotheses</li> <li>• 2: performs adequate tests to assess key hypotheses but does not discuss the results of the tests</li> <li>• 3: performs adequate tests to assess key hypotheses and critically assesses and comments on the results of the tests</li> </ul> <p><u>0-16: Regression</u> 0: no attempt</p> <p>0-2: Modelling</p> <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 1: only one regression model is examined</li> <li>• 2: several regression models are examined</li> </ul> <p>0-2: Discussion of the output of the regression models</p> <ul style="list-style-type: none"> <li>• 0: no discussion</li> <li>• 1: limited discussion of models' output or discussion of only one model</li> <li>• 2: detailed discussion of models' output</li> </ul> <p>0-2: Outliers and influential observations</p> <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 1: outliers and influential observations are identified</li> <li>• 2: outliers and influential observations are identified, and appropriate visualizations are provided</li> </ul> <p>0-4: Diagnostics</p> <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 1: diagnostics are performed and assessed for only one model</li> <li>• 2: diagnostics are performed and assessed for only one model and appropriate visualizations are provided</li> <li>• 3: diagnostics are performed and assessed for all models examined</li> </ul>

Description	Range	Break down
		<ul style="list-style-type: none"> <li>• 4: diagnostics are performed and assessed for all models examined and appropriate visualizations are provided</li> </ul> <p>0-3: Transformations</p> <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 0-1: transformations of independent variables are justified and performed</li> <li>• 0-1: transformations of the dependent variable are justified and performed</li> <li>• 0-1: appropriate visualizations are provided</li> </ul> <p>0-3: Progress of regression modelling</p> <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 0-1: the various models examined are compared</li> <li>• 0-1: the steps to result to the final model are explained and justified</li> <li>• 0-1: provision of the resulting regression model equation</li> </ul> <p><u>Requirements for all generated plots</u></p> <ul style="list-style-type: none"> <li>• Appropriate labelling of axes</li> <li>• Appropriate plot titles</li> <li>• Data labels where appropriate</li> </ul>
R code	0-11	<p><u>0-8 R code implementation</u></p> <p>0: no attempt OR code does not compile in a clean workspace</p> <p>0-1: code compiles in a clean workspace and reproduces the output presented in the report</p> <p>0-1: Consistent style with code</p> <p>0-2: Code reproduces the output of the hypothesis testing section</p> <p>0-2: Code reproduces the output of regression modelling section</p> <p>0-2: Code reproduces the output of the diagnostics section</p> <p><u>0-3 Commenting</u></p> <p>0: does not comment on the R code</p> <p>1: provides limited comments, inadequate to explain the code</p>

Description	Range	Break down
		2: provides adequate comments in the code 3: provides detailed comments, clearly explaining all the steps of the code and how the commands correlate with the statistical methods described in the report
Presentation	0-15	0: no presentation  0-1: Introduction (description of the dataset, aim and objectives of the study)  0-5: Descriptive statistical analysis <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 1: explanation of descriptive analysis methods employed</li> <li>• 0-2: presentation of descriptive analysis results (none / limited / adequate)</li> <li>• 0-2: visualizations presented (none / limited / adequate)</li> </ul> 0-5: Inferential statistical analysis <ul style="list-style-type: none"> <li>• 0: no attempt</li> <li>• 1: explanation of inferential analysis methods employed</li> <li>• 0-2: presentation of inferential analysis results (none / limited / adequate)</li> <li>• 0-2: visualizations presented (none / limited / adequate)</li> </ul> 0-1: Conclusions <ul style="list-style-type: none"> <li>• 0: no conclusions are presented</li> <li>• 1: conclusions are presented</li> </ul> 0-1: Presentation quality <ul style="list-style-type: none"> <li>• 0: poor quality presentation</li> <li>• 1: high quality presentation</li> </ul> 0-2: Answering of questions made by the class after the presentation <ul style="list-style-type: none"> <li>• 0: fails to answer any questions</li> <li>• 1: answers some questions made</li> <li>• 2: answers all questions made</li> </ul>

Make sure that your R code contains all required commands so that the code will run on a clean workspace without producing an error, and replicating the results described in your report. If your R code replication produces an error, marks will be deducted from your R code score.

Note that since Assignment 2 is a continuation of Assignment 1, you will be required to run part of your Assignment 1's R code again in order to continue with the processing for Assignment 2.

### **Explanations with regard to Statistical Analysis**

The statistical analysis must include

- Hypothesis testing, e.g. t-test (independent or paired accordingly), correlation test, etc.
- Multiple Linear Regression
- Regression diagnostics
- Transformation of the dependent and/or independent variables

### **In general**

- The study objectives should be clearly stated.
- The specific question your analysis aims to address should be clearly formulated.
- All output of the analysis must be reported.
- Clearly indicate the independent and dependent variables.
- Provide details of the proposed analysis.
- Defend on any necessary relabel of variables or data transformation.
- Defend on the number of significant digits used to reflect on precision.
- Figures and tables should clearly display the data and be able to stand alone. That is all information necessary for interpretation should be included within the figure/table and legend.
- Sufficient detail regarding hypothesis testing should be provided.
- Specify and defend on whether one or two sided statistical tests are performed.
- The presence of outliers should be noted. State your approach to handle them.
- Report exact p-values.

### **Submission of assignment work**

- Anonymous marking is being used. You may include your University ID number ("G2...") on the work. Apart from this, avoid doing anything that would allow you to be identified from your work.
- *Keep a complete copy of the work you hand in.*
- Avoid submitting work at the last minute, but if there is a technical problem uploading to Blackboard, email the zip file to me before the deadline and upload the work when Blackboard is available.

### **Extenuating circumstances, extensions and late work**

Except where an extension of the hand-in deadline date has been approved (see [https://www.uclan.ac.uk/students/study/examinations\\_and\\_awards/extenuating\\_circumst](https://www.uclan.ac.uk/students/study/examinations_and_awards/extenuating_circumst)

[ances.php](#)), work that is handed in up to 5 days late will be capped to 50%. After this, it will receive a mark of 0%:

### **Cheating**

The consequences of cheating in assessments are serious. Cheating is using or attempting to use unfair means to enhance performance. This includes plagiarism (presenting someone else's work as if it was your own), collusion (working with others on an individual assignment), taking prohibited material into examinations and allowing other students to access your work. Make sure that you do not give someone the opportunity to steal your work (e.g. *by asking them to print it out for you*). We tell students about cheating both during induction and in your student handbook, but if you have any doubt about what cheating is or how to reference material properly, please ask a tutor. We recommend that you use the Harvard system for referencing.

The University operates an electronic plagiarism detection service where your work may be uploaded, stored and cross-referenced against other material. The software searches the World Wide Web and extensive databases of reference material to identify duplication.

For more information about plagiarism, please see the University Academic Regulations and the Assessment Handbook ([http://www.uclan.ac.uk/aqasu/academic\\_regulations.php](http://www.uclan.ac.uk/aqasu/academic_regulations.php)). See the Student Union website: <http://www.uclansu.co.uk/academicmatters/unfairmeans>

### **Reassessment and Revision**

Reassessment in written examinations and coursework is at the discretion of the Course Assessment Board and is dealt with in accordance with University policy and procedures.